Spatial Analysis of Fatal and Injury Crashes in Flanders, Belgium; Application of Geographically Weighted Regression Technique

Ali Pirdavani, Tom Brijs,
Tom Bellemans and Geert Wets[1]

Transportation Research Institute (IMOB)
Hasselt University
Wetenschapspark – Building 5
B-3590 Diepenbeek, Belgium

Tel: +32(0)11 26 91{39, 55, 27, 47, 58}
Fax: +32(0)11 26 91 99

Email: {ali.pirdavani, tom.brijs, tom.bellemans, geert.wets}@uhasselt.be

Total number of words: 7563 (3 Tables and 3 Figures included)
Date of 1st submission: 24/07/2012
Date of 2nd submission: 15/11/2012

[1] Corresponding author

ABSTRACT

Generalized Linear Models (GLMs) are the most widely used models utilized in crash prediction studies. These models illustrate the relationships between the dependent and explanatory variables by estimating fixed global estimates. Since the crash occurrences are often spatially heterogeneous and are affected by many spatial variables, the existence of spatial correlation in the data is examined by means of calculating Moran's *I* measures for dependent and explanatory variables. The results indicate the necessity of considering the spatial correlation when developing crash prediction models. The main objective of this research is to develop different Zonal Crash Prediction Models (ZCPMs) within the Geographically Weighted Generalized Linear Models (GWGLM) framework in order to explore the spatial variations in association between Number of Injury Crashes (NOICs) (including fatal, severely and slightly injury crashes) and other explanatory variables. Different exposure, network and socio-demographic variables of 2200 Traffic Analysis Zones (TAZs) are considered as predictors of crashes in the study area, Flanders, Belgium. To this end, an activity-based transportation model framework is applied to produce exposure measurements while the network and socio-demographic variables are collected from other sources. Crash data used in this study consist of recorded crashes between 2004 and 2007. GWGLMs are developed using a Poisson error distribution and are often referred to as Geographically Weighted Poisson Regression (GWPR) models. Moreover, the performances of developed GWPR models are compared with their corresponding GLMs. The results show that GWPR models outperform the GLM models; this is due to the capability of GWPR models in capturing the spatial heterogeneity of crashes.

1. INTRODUCTION

For many years, researchers have attempted to investigate the negative impacts of growing travel demand on traffic safety by predicting the number of crashes based on the patterns they have learnt from crashes that occurred in the past. This should lead to providing a predictive tool which is capable of evaluating road safety at the planning-level. Dealing with traffic safety at the planning-level requires the ability to integrate Travel Demand Management (TDM) policies into a crash predicting context. TDM policies are usually performed at a more aggregate level than just on the level of an individual intersection or road section. Thus, the impact of adopting a TDM strategy on transportation or traffic safety should be evaluated at a higher level rather than merely the local consequences. Application of Crash Prediction Models (CPMs) at a macro level like Traffic Analysis Zone (TAZ) leads to a type of prediction models commonly referred to as Zonal Crash Prediction Models (ZCPMs). ZCPMs have been initially introduced by Levine et al. based on Linear Regression models (*1*). However, the most common modeling framework for ZCPMs is the Generalized Linear Modeling (GLM) framework (*2–19*). Within a GLM framework, fixed coefficient estimates explain the association between the dependent variable and the explanatory variables. In other words, a single model tries to fit the observed data for all locations (TAZs) similarly. Expectedly, different spatial variation may be observed for different explanatory variables especially where the study area is relatively large. Neglecting this spatial variation may deteriorate the predictive power of ZCPMs.

Spatial variation is known to be an important aspect of traffic safety analysis and in particularly crash prediction modeling. Inclusion of spatial variation in traffic safety studies has been reported by many researchers. In one of the earliest studies, the spatial relationship between activities which generate trips and motor vehicle accidents was studied and applied to the City and County of Honolulu (*20*). Different spatial patterns for different variables such as population, employment and road characteristics were identified. LaScala et al. (*21*) found that significant spatial relationships exist between specific environmental and demographic characteristics of the City and County of San Francisco and pedestrian injury crashes. Flahaut et al. (*22*) presented different methods for identifying and delimiting accidents black-zones. This was an application of spatial correlation in defining accident black-zones which share similar characteristics. A similar study was carried out by Moons et al. (*23*) where the structure of the underlying road network is taken into account by applying Moran's *I* to identify crash hot zones. In another study by Flahaut (*24*), it was indicated that spatial autocorrelation should be integrated in the modeling process if spatial data are being studied. He concluded that spatial models in comparison to non-spatial models, do not overestimate the significance of explanatory variables; thus, spatial variation should be considered to analyze spatial data. Geurts et al. (*25*) investigated the clustering phenomenon in road accidents. This was an application of spatial analysis in traffic safety that aims to analyze the characteristics of specific zones on which more accident occur. Spatial correlation was found to be significant in injury crashes in a study conducted for the State of Pennsylvania at the county level (*9*). Aguero-Valverde and Jovanis (*26*) further explored the effect of spatial correlation in models of road crash frequency at the segment level. The results of their study highlighted the importance of including spatial correlation in road crash modeling studies. The models with spatial correlation show significantly better fit compared to the Poisson lognormal models. The existence of clusters in the spatial arrangement of pedestrian crashes was reported by (*27*). They supported their conclusions by computing Moran's *I* value and presenting the Local Indicators of Spatial Association (LISA) significance map of crashes. Huang et al. (*28*) performed a county-level road safety analysis for the state of Florida. They reported that significant spatial correlations in crash occurrence were identified across adjacent counties.

This spatial variation which is often referred to as "spatial non-stationarity" is overlooked by the GLMs. Following such a modeling approach ends up with a set of fixed global variable estimates which are the same for different TAZs; however, it is possible that an explanatory variable which is found to be a significant predictor of crashes in some TAZs might not be a powerful predictor in other TAZs. There are different spatial modeling techniques that have been applied by many researchers in the crash prediction field. Auto-logistic models, Conditional Auto-regression (CAR) models, Simultaneous Auto-regression (SAR) models, spatial error models (SEM), Generalized Estimating Equation (GEE) models, Full-Bayesian Spatial models, Bayesian Poisson-lognormal models are some of the most employed techniques to conduct spatial modeling in traffic safety (*9*), (*20*), (*24*), (*26*), (*28–34*). The output of these models are still fixed variable estimates for all locations, however the spatial variation is taken into account. Another solution for taking the spatial variation into account is developing a set of local models, so called Geographically Weighted Regression (GWR) models (*35*). These models rely on the calibration of multiple regression models for different geographical entities. The GWR approach has mainly been followed in health, economic and urban studies. Also a few studies have been carried out in the transportation field using this technique (*36–41*). In traffic safety, Hadayeghi et al. (*3*) developed GWR models to investigate spatial variations in the model relationships. The results of the GWR models indicated an improvement in model predictability by means of an increased R^2. In another study (*42*), bicycle crashes were studied in Buffalo, New York. Density of development, physical road characteristics, socioeconomic and demographic variables were the selected explanatory variables. Given the spatial nature of these variables, a GWR model was developed and showed a better performance compared with the conventional model. An inter-province difference in traffic accidents in Turkey was studied by Erdogan (*43*). Different spatial autocorrelation analyses were performed to see whether the accidents are clustered or not. Since the results of these analyses indicated non-stationarity in the data, a GWR model was developed. They also showed that the GWR model performs better than the Ordinary Least Square (OLS) model.

The GWR technique can be adapted to GLM models and form Geographically Weighted Generalized Linear Models (GWGLMs) (*35*). GWGLMs are able to serve the count data (such as number of crashes) while simultaneously accounting for the spatial non-stationarity. Hadayeghi et al. (*11*) used the GWR technique in conjunction with the GLM framework using the Poisson error distribution. They developed different Geographically Weighted Poisson Regression (GWPR) models to associate the relationship between crashes and a number of predictors. The results of the comparisons between GLMs and GWPR models revealed that the GWPR models outperform the GLMs since they are capable of capturing spatially dependent relationships.

The first objective of this paper is to examine the existence of spatial correlation in the dependent and other explanatory variables available in the data. The main objective of this study is then to develop different ZCPMs within the GWGLM framework in order to explore the spatial variations in association between crashes and other explanatory variables. Moreover, the performance of GWGLMs will be compared with the GLMs developed in an earlier study (*44*). In this study GWGLMs are developed using a Poisson error distribution; henceforth, we refer to these models as GWPR models.

2. METHODOLOGY

2.1. Data Preparation

The required information to construct the prediction models consists of exposure, network and socio-demographic data accompanied with the crash data. These data should be collected for the whole study area and also be aggregated to the zonal level. The study area in this research is the Dutch speaking region in northern Belgium, Flanders.

Exposure is an important determinant of traffic safety. Therefore, it is essential to have the exposure metrics as accurately as possible. To this end, the FEATHERS (Forecasting Evolutionary Activity-Travel of Households and their Environmental RepercussionS) activity-based transportation model is applied. The FEATHERS framework (*45*) was developed in order to facilitate the development of activity-based models for transportation demand in Flanders, Belgium. The real-life representation of Flanders is embedded in an agent-based simulation model which consists of over 6 million agents, each agent representing one member of the Flemish population. A sequence of 26 decision trees is used in the scheduling process and decisions are based on a number of attributes of the individual (e.g. age, gender), of the household (e.g. number of cars) and of the geographical zone (e.g. population density, number of shops). For each individual with its specific attributes, the model simulates whether an activity (e.g. shopping, working, and etc.) is going to be carried out or not. Subsequently, the location, transport mode and duration of the activity are determined, taking into account the attributes of the individual (*46*). As such, the FEATHERS activity-based model can provide the exposure measure by means of time-of-day dependent Origin-Destination (OD) matrices for all traffic modes (i.e. Number of Trips (NOTs)). Assigning the OD matrices of car trips to the Flemish road network provides other exposure variables like Vehicle Kilometers Traveled (VKT) and Vehicle Hours Traveled (VHT). These network level exposure measures are then aggregated to the zonal level comprising of 2200 TAZs. In addition, for each TAZ a set of socio-demographic and network variables were derived. The crash data used in this study consist of a geo-coded set of injury crashes (including fatal, severely injured and slightly injured crashes) that have occurred during the period 2004 to 2007. Table 1 shows a list of variables, together with their definition and descriptive statistics, which have been used in developing the models presented in this paper.

2.2. Motivation for Conducting Spatial Analysis

Previous research has indicated that there might be significant spatial correlations in crash occurrence across different locations TAZs; e.g. (*11*), (*27*), (*28*), (*34*), (*43*). Therefore, it is essential to check for the existence of spatial correlation of dependent and explanatory variables. This can be carried out by means of different statistical tests such as Moran's autocorrelation coefficient commonly referred to as Moran's *I*. The results of the analysis indicate the necessity of considering this spatial correlation since the spatial status of all variables are found to be non-stationary.

TABLE 1 List of Explanatory Variables for the ZCPMs with Their Definition and Descriptive Statistics

	Variable	Definition	Average	Min	Max	SD[a]
	Crash	total NOICs observed in a TAZ	36.03	0	326	41.58
Exposure variables	Number of Trips	average daily number of trips originating/destined from/to a TAZ	2765.8	0	18111.4	2869.8
	Total Flow	Average Annual Daily Traffic (AADT) in a TAZ (vehicle)	96414.5	70.9	4423325	181695
	VHT	total daily vehicle hours traveled in a TAZ	608.26	1.50	9998.6	930.29
	VKT	total daily vehicle kilometers traveled in a TAZ	52533.8	84.06	985192	90715.2
	Motorway Flow	AADT of motorways in a TAZ (vehicle)	37724.96	0	3881777	146757.5
	Motorway VHT	total daily vehicle hours traveled on motorways in a TAZ	260.52	0	9762.5	832.97
	Motorway VKT	total daily vehicle kilometers traveled on motorways in a TAZ	27471.82	0	946152.8	84669.53
	Other Roads Flow	AADT of other roads in a TAZ (vehicle)	58690.29	0	734152.5	73632.5
	Other Roads VHT	total daily vehicle hours traveled on other roads in a TAZ	348.51	0	3777.69	358.76
	Other Roads VKT	total daily vehicle kilometers traveled on other roads in a TAZ	26662.85	0	303237.6	28133.04
	V/C	average volume to capacity in a TAZ	0.0478	0	0.5697	0.0422
Network variables	Speed	average speed limit in a TAZ (km/hr)	69.4	31	120	10.91
	Capacity	hourly average capacity of links in a TAZ	1790.1	1200	7348.1	554.6
	Area	total surface area of a TAZ in square kilometers	6.09	0.09	45.22	4.78
	No. of Links	number of links in a TAZ	39.27	1	230	30.46
	Link Length	total length of the links in a TAZ (km)	15.86	0.39	87.95	10.79
	Link Density	link length per square kilometers in a TAZ	3.37	0.03	20.44	2.41
	Intersection	total number of intersections in a TAZ	5.8	0	40	5.9
	Intersection Density	number of intersection per square kilometers	1.76	0	50.63	3.39
	Motorway	presence of motorway in a TAZ describes as below: "No" represented by 0 "Yes" represented by 1	0	0	1	-[b]
	Urban	Is the TAZ in an urban area? "No" represented by 0 "Yes" represented by 1	0	0	1	-
	Suburban	Is the TAZ in a suburban area? "No" represented by 0 "Yes" represented by 1	0	0	1	-
Socio-demographic variables	Driving License	average driving license ownership in a TAZ describes as percentage	81.1	0	100	3.5
	Income Level	average income of residents in a TAZ describes as below: "Monthly salary less than 2249 Euro" represented by 0 "Monthly salary more than 2250 Euro" represented by 1	1	0	1	-
	Work Status	average work status of the residents in a TAZ describes as below: "Don't work" represented by 0 "Work" represented by 1	1	0	1	-
	Population	total number of inhabitants in a TAZ	2614.52	0	15803	2582.6
	Population Density	population per square kilometers	774.14	0	14567.4	1398.4

a: Standard deviation b: Data not applicable.

2.3. Model Construction

Generalized Linear Model

Reviewing the literature for different model forms showed that the following model has been widely used in different studies (*7*), (*14*), (*17*), (*18*):

$$E(C) = \beta_0 \times (Exposure)^{\beta_1} \times e^{\sum_{i=2}^{n} \beta_i x_i} \quad (1)$$

Where $E(C)$ is the expected crash frequency, β_0 and β_i are model parameters, $Exposure$ is the exposure variable (e.g. VHT, VKT or NOTs) and x_i's are the other explanatory variables.

Logarithmic transformation of equation (5) when considering only one exposure variable yields:

$$ln[E(C)] = ln(\beta_0) + \beta_1 ln(Exposure) + \beta_2 x_2 + \beta_3 x_3 + \ ... + \beta_n x_n \quad (2)$$

Geographically Weighted Generalized Linear models

The Geographically Weighted form of Equation (6) would be:

$$ln[E(C)(\boldsymbol{l}_i)] = ln(\beta_0(\boldsymbol{l}_i)) + \beta_1(\boldsymbol{l}_i) ln(Exposure) + \beta_2(\boldsymbol{l}_i) x_2 + \cdots + \beta_n(\boldsymbol{l}_i) x_n \quad (3)$$

The output of these models will be different location-specific estimates for each case (here each TAZ). All variable estimates are functions of each location (here the centroid of each TAZ), $\boldsymbol{l}_{\mathrm{i}} = (x_{\mathrm{i}}, y_{\mathrm{i}})$ representing the x and y coordinates of the i^{th} TAZ. The main purpose of developing geographically weighted models is that these models allow the estimates to vary where different spatial correlation among the explanatory variables exists. If the aim is estimating parameters for a model at a specific location, expectedly the locations nearby this location have a greater impact on this estimation compared with the locations which are far from it. This impact can be expressed by a weighting function. This weighting function is conditioned on the location $\boldsymbol{l}_i$ and changes for each location (*35*). The weights are derived from a weighting scheme which is commonly referred to as a kernel. There are two kernels which are frequently used to generate the weighting scheme; the Gaussian and the bi-square functions which can be formulated as follows:

Gaussian function: $\mathrm{W}_{\mathrm{ij}} = e^{-0.5(\frac{d_{\mathrm{ij}}}{\mathrm{b}})^2}$ (4)

bi-square function: $\mathrm{W}_{\mathrm{ij}} = \begin{cases} (1 - \left(\frac{d_{\mathrm{ij}}}{\mathrm{b}}\right)^2)^2 & \text{if } \ d_{\mathrm{ij}} < b \\ 0 & \text{otherwise} \end{cases}$ (5)

Where W_{ij} represents the measure of contribution of location j when calibrating the model for location i, d_{ij} is the distance between locations i and j and b is the bandwidth (*35*). It is reported in the literature (*11*), (*47*) that selection of the kernel function and accordingly the bandwidth is very critical as the model might be very sensitive to this selection. However, Fotheringham et al. (*35*) indicated that regarding the fit of the model, the choice of a bandwidth is more important than the shape of the kernel. As a rule of thumb, when the sample locations are commonly positioned across the study area, then a kernel with a fixed bandwidth is a suitable choice for modeling. On the contrary, when the sample locations are clustered in the study area, it is generally better to apply an adaptive kernel; i.e., having larger bandwidth where sample locations are sparser and applying smaller bandwidth for denser sample locations. Adaptive bandwidth will be displayed as a quantile of the number of adjacent locations (TAZs) which will influence the weighting function (e.g. in Table 2 and for Model#4, the bandwidth value is 0.03369; this means that 3.369% of the adjacent TAZs, 74 TAZs out of 2200 TAZs, should be selected to calculate the weighting function for each TAZ).

Despite the fact that parameters estimation depends on the weighting function, selecting an appropriate bandwidth is a more crucial task. There are different approaches that can be used in bandwidth selection. Cross-validation (CV) is a technique in which the optimal bandwidth size is determined by minimizing the CV score which is formulated as follows:

$$CV = \sum_{i=0}^{n} (\mathrm{y_i} - \hat{\mathrm{y}}_{\neq \mathrm{i}})^2 \tag{6}$$

Where n is the number of TAZs and $\hat{\mathrm{y}}_{\neq \mathrm{i}}$ is the fitted value of $\mathrm{y_i}$ when the i$^{\text{th}}$ case is left out during the calibration process.

Another method to derive the bandwidth which provides a trade-off between Goodness-of-fit and degrees of freedom is minimizing the Akaike Information Criterion (AIC) (*35*). It is reported by (*48*) that in the case of local regression, given the fact that the degrees-of-freedom are likely to be small, including a small sample bias adjustment in the AIC definition is recommended. This will lead to a corrected AIC often referred to as AICc. The formulations of AIC and AICc are as follows:

$$AIC = D(b) + 2K(b) \tag{7}$$

and

$$AICc = AIC + 2\frac{K(b)(K(b)+1)}{n - K(b) - 1} \tag{8}$$

Where D and K are respectively the deviance and the effective number of parameters in the model with bandwidth b and n denotes the number of TAZs.

In this study, both the CV and AICc methods were applied to determine the most appropriate bandwidth. The results reveal that in case of applying the AICc method, the optimum derived bandwidths are very close to each other no matter which kernel function is used. The computed bandwidths following the CV approach are slightly different than the ones derived by the AICc approach. Since the model selection is based on the minimum AICc values, only the bandwidths derived by the AICc approach will be used in model development.

Model development and spatial analysis are carried out using the statistical software package R (*49*) and GWPR models are developed using a SAS macro (*50*) .

3. DISCUSSION ON MODEL RESULTS

3.1. Finding the Best Fitted Model

A common rule-of-thumb in the use of AICc is that if the difference in AICc values between two models is more than 2, there is a substantial difference in the performance of the two models (*48*). As can be seen in Table 2, Model#4 outperforms all other models by means of having the minimum AICc value which is far lower than the AICc values of all other models. Model#4 is fitted using an adaptive bandwidth and a Gaussian kernel function for the weighing function. It can be concluded that for the given data, utilizing adaptive bandwidth and the Gaussian kernel function will result in the best model fit. Therefore, this combination will be used to fit different models by which we aim to compare the performance of GWPR models against GLM models.

TABLE 2 Comparison between GLM and GWPR Models

	Model #1	Model #2	Model #3	Model #4
Coefficients	Estimates	Estimates	Estimates	Estimates
(Intercept)	-4.141e+00	-2.886e+00	-6.32, -1.156 (-4.39,-3.64,-2.84) [a]	-5.215, -0.1736 (-3.26,-2.75,-1.92)
ln(Number of Trips)	4.520e-01	4.676e-01	0.1375, 0.7652 (0.37,0.48,0.59)	0.1471, 0.7665 (0.36,0.462,0.57)
ln(Motorways VKT)	7.744e-03	-	-0.027, 0,0217 (-0.006,0.001,0.0098)	-
ln(Other Roads VKT)	3.132e-01	-	0.1298, 0.4188 (0.21,0.25,0.303)	-
ln(Motorways VHT)	-	7.717e-03	-	-0.0385, 0.0366 (-0.013,-0.002,0.011)
ln(Other Roads VHT)	-	3.040e-01	-	0.1269, 0.4684 (0.229,0.27,0.343)
Capacity	3.894e-04	4.220e-04	-1.5e-4, 7.61e-4 (1.7e-4,3.1e-4,4.3e-4)	-8.8e-5, 7.1e-4 (2.1e-4,3.3e-4,4.4e-4)
Intersection	2.888e-02	2.844e-02	0.005, 0.052 (0.02, 0.026,0.031)	-0.0042, 0.053 (0.02,0.026,0.03)
Income level	-1.071e-01	-1.056e-01	-0.526, 0.498 (-0.195,-0.072,0.01)	-0.5875, 0.5099 (-0.19,-0.064,0.023)
Urban	3.520e-01	2.287e-01	-0.137,0.783 (0.291,0.394,0.56)	-0.2487, 0.6552 (0.204,0.33,0.48)
Suburban	9.095e-02	5.712e-02	-0.102, 0.384 (0.07,0.145,0.233)	-0.1299, 0.376 (0.063,0.139,0.215)
Population	2.293e-05	2.340e-05	-5.4e-5, 9.23e-5 (9.6e-7,2.2e-5,3.7e-5)	-5.6e-5, 9.26e-5 (-2e-6,2.1e-5,3.5e-5)
Bandwidth	-	-	0.03371	0.03369
AICc	16918	16921	10713	10605
MSPE	489.74	482.41	238.27	234.82
PCC	0.869	0.871	0.929	0.931

a: minimum, maximum, (1st quartile, median, 3rd quartile) of the parameter estimates.

Model #1: GLM Negative Binomial model using VKT and NOT
Model #2: GLM Negative Binomial model using VHT and 0OT
Model #3: GWPR model with adaptive bandwidth and Gaussian kernel using VKT and NOT
Model #4: GWPR model with adaptive bandwidth and Gaussian kernel using VHT and NOT

Comparable with our previous research (*44*) in which different GLM models were developed, similar GWPR models are constructed to evaluate the benefits of accounting for the spatial autocorrelation. The GWPR models and their corresponding GLM models are summarized and their performances together with their goodness-of-fit measures are presented in

Tables 2. There are a number of measures that have been used in comparative analysis between different models; e.g. AICc, MSPE and PCC (*14*). Comparing AICc, PCC and MSPE measures in Table 2 shows that GWPR models outperform the GLM models.

3.2. Further Investigation on the Selected Model

As stated earlier, the results of the GWPR models are presented as sets of locally estimated coefficients often referred to as 5-number summaries (i.e. minimum, 1st quartile, median, 3rd quartile and maximum of coefficient estimates of all local models). Unlike spatially stationary models (e.g. GLM models) which have a single estimate for each variable, variable estimates for GWPR models vary across the space and sometimes have different and unexpected signs. Unlike some other studies (*11*) which report on this trend to happen for their most significant variables, in our study all of the most significant variables have similar signs in line with our expectations. “ln(Number of Trips)” and “ln(Other Roads VHT)” as the most significant variables always have positive signs for all local estimates. However the signs of other coefficients are not always the same. To have a better view on these differences, local variable estimates are depicted in Figure 1. This issue which is often referred to as “the problem with counterintuitive signs” has already been reported in many studies (*11*), (*38*), (*47*). One explanation for this problem would be the existence of multicollinearity among some variables for some locations. It is quite possible that some variables at some locations are locally correlated while no global multicollinearity observed among the explanatory variables.

Another reason could be due to the basis of calibrating GWPR models. Presumably for some locations, some variables might not be significant variables; therefore, it is possible that the local models produce some unexpected variable signs for those insignificant variables. The latter reason can be easily investigated by mapping local p-values. Figure 2 depicts p-values for all explanatory variables of Model#4. In this figure, significant variables at any location are colored in green while insignificant variables are depicted in red. By comparing Figures 1 and 2 it can be concluded that the p-values for all of the locations with unexpected coefficient signs are insignificant at the 95% confidence level. For instance, the variable “Urban” is expected to have a positive association with the crash frequency (*28*), (*44*). As can be seen from Figure 1, only a few TAZs show negative association with the NOICs (i.e. TAZs colored in green). When comparing this figure with the corresponding map in Figure 2, it is evident that in these TAZs, “Urban” is not a significant predictor. This is similar for other explanatory variables where the TAZs with unexpected variable signs are always the TAZs where variables are insignificant predictors.

Generally, the GWPR models outperform the GLM models because of their capability in capturing spatial heterogeneity. As can be seen from Figure 3, observed and predicted NOICs are having almost the same pattern. This is an indication of how well these models are able to fit the observed data.

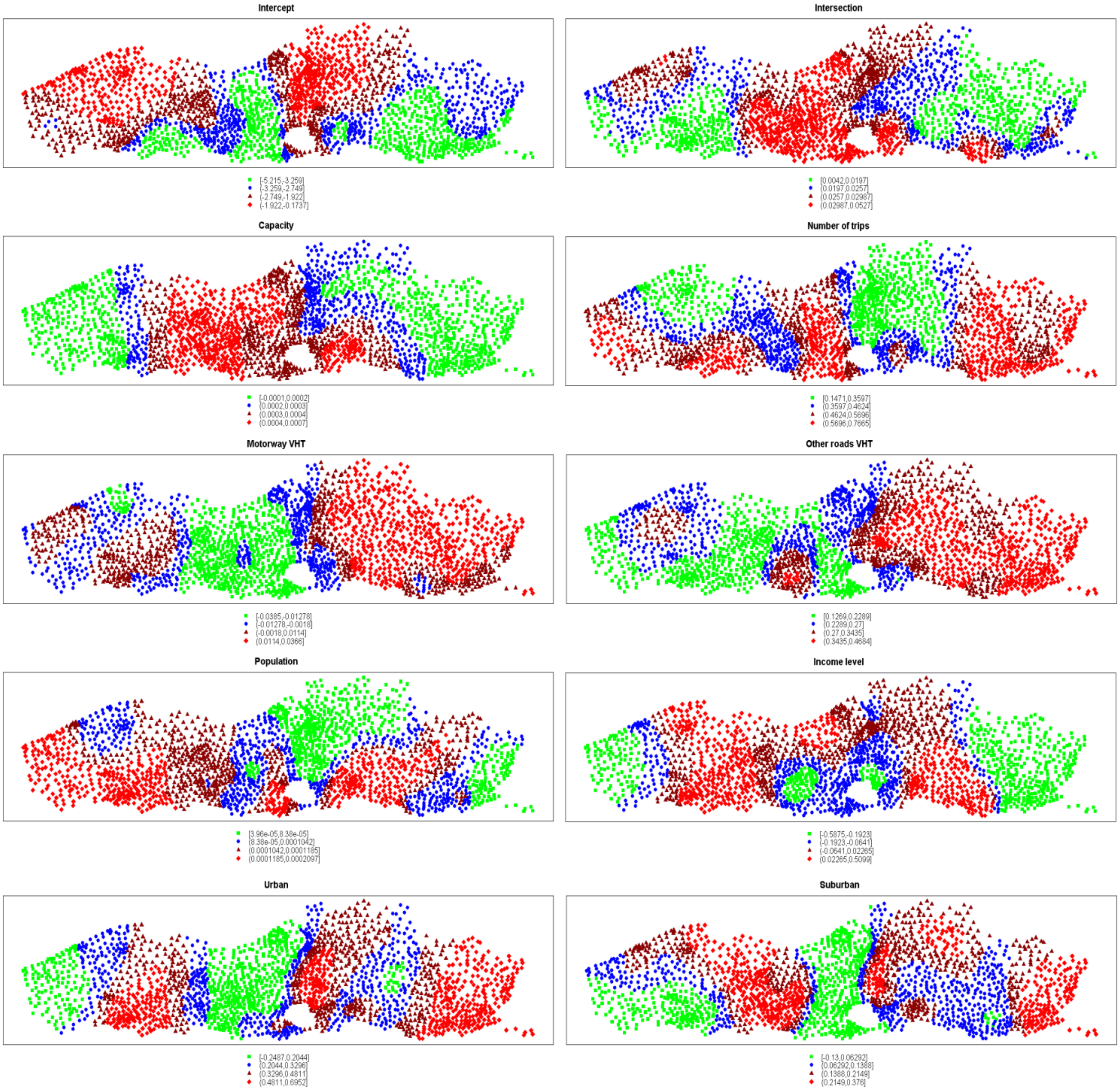

FIGURE 1 Graphical representation of local variable estimates for Model#4.

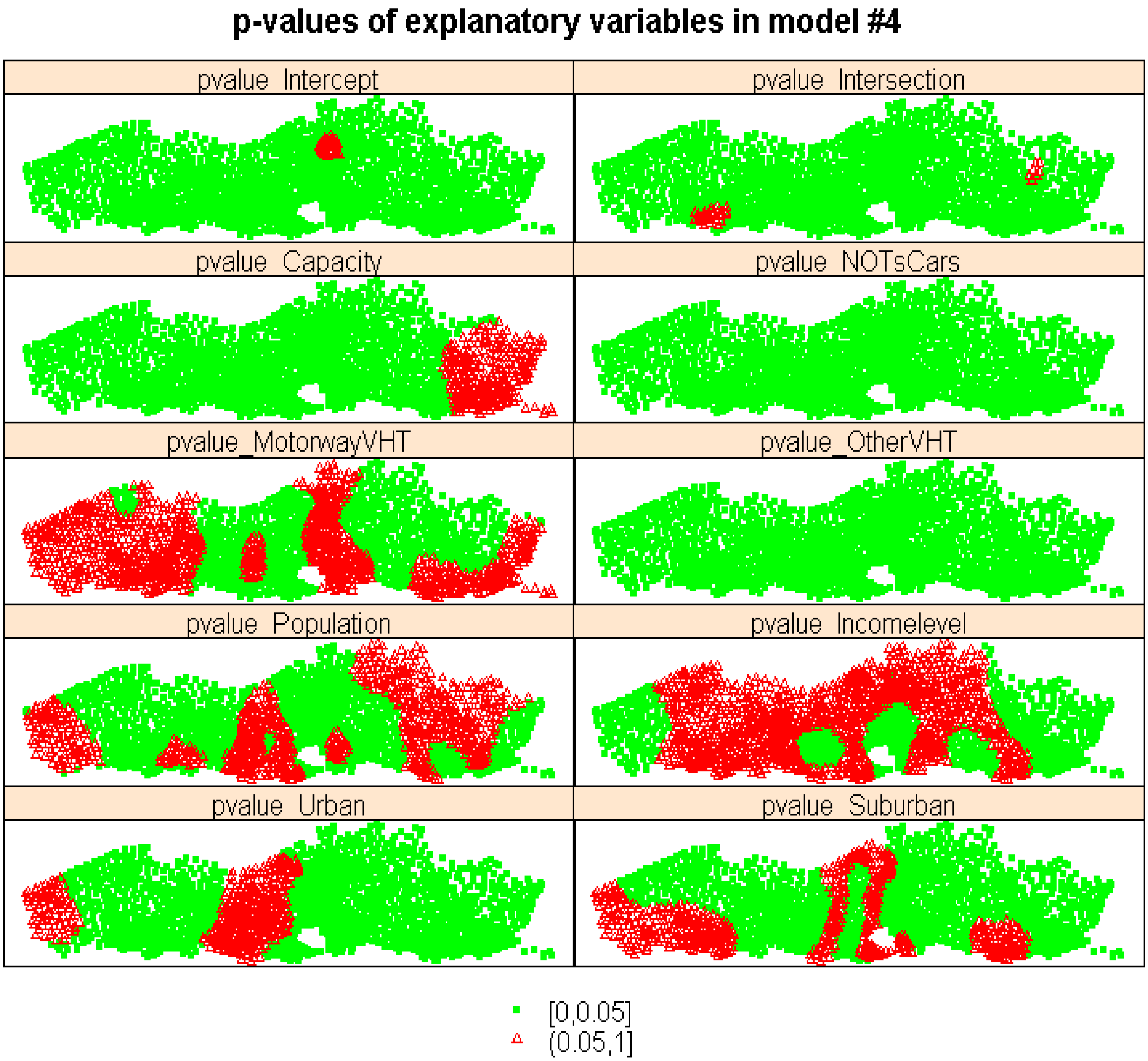

FIGURE 2 Graphical representations of p-values of all explanatory variables for Model#4.

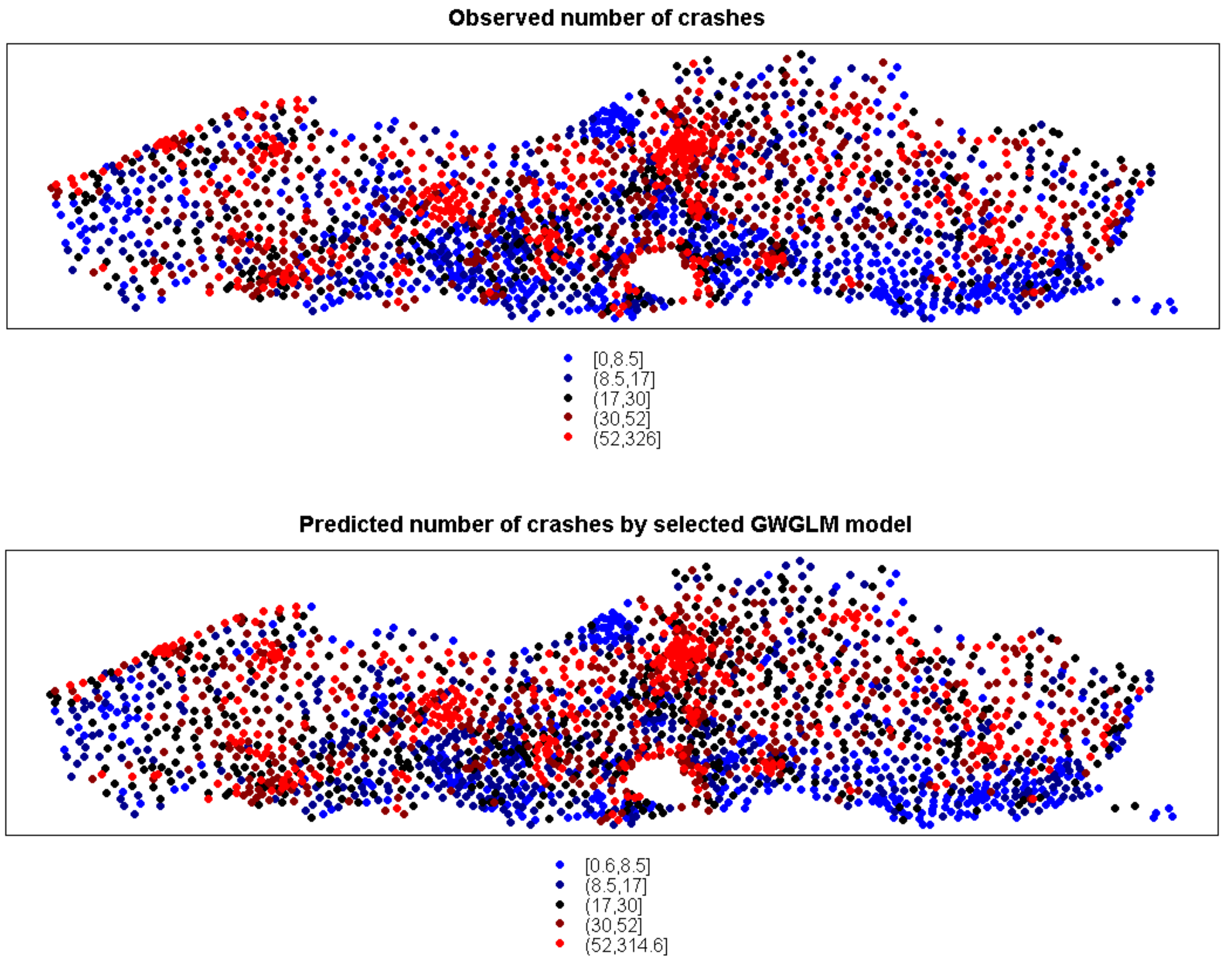

FIGURE 3 Observed and predicted (results from Model#4) NOICs.

4. VALIDATION

Strong dependencies among the local coefficient estimates imply the fact that coefficients are not uniquely defined and as such, any convincing interpretation cannot be derived (*51*). Due to the greater complexities of the GWR estimation procedure that conceivably causes interrelationships among the local estimates, it is essential to check for multicollinearity among local coefficient estimates. There are frequently used exploratory tools available to discover possible multicollinearity, such as bivariate scatter plots or bivariate correlation coefficients, however, a more statistically oriented measure that adopts a simultaneous view to identify multicollinearity is variance inflation factor (VIF). The VIF quantifies the severity of multicollinearity. It provides an index that measures how much the variance of an estimated regression coefficient is increased because of collinearity. Analyzing the magnitude of multicollinearity is carried out by considering the size of the VIF. As a common rule of thumb, 10 is defined (*52*) as a cut off value meaning that if the VIF is higher than 10 then multicollinearity is high. VIF values among local coefficient estimates of models are shown in Table 3. These results suggest that multicollinearity among local coefficient estimates is not a problem in any of the developed models.

TABLE 3 VIF Among Local Coefficient Estimates of Model#4

Coefficients	VIF value
NOTsCars	1.550481
Motorways VHT	4.309746
Other Roads VHT	2.719767
Incomelevel	2.124447
Capacity	3.42898
Intersection	3.827387
Urban	2.347726
Suburban	1.759584
Population	2.035026

Due to the nature of GWR models which are location specific models, validation cannot be accomplished by means of conventional methods (e.g. k-fold cross validation). Unlike traditional regression modeling in which a general model is fitted on training dataset and validated on a test dataset, GWR models are a series of local models, therefore, the concept of training and testing cannot be applied in the context of GWR models. However, a new framework is proposed in this research by which sensitivity of the predictability power of fitted models is checked. To this end, the whole dataset is randomly divided into 10 segments. In each round of model fitting one segment is left out, therefore, there will be 9 different models fitted for each single data point (here TAZ). Each of these models are developed by using the derived information from the neighboring TAZs. In this case, neighboring TAZs are changed in each round of model fitting for each TAZ. Robustness of the prediction models can be confirmed by checking the variability of predictions derived from 9 different models that are fitted for each TAZ. In case of having an acceptable low variation in predictions, it could be concluded that models are not sensitive to presence/absence of specific vicinity TAZs. Moreover, a low variation in predictions further confirms presence of spatial correlation and the right choice of bandwidth, meaning that missing information of left out TAZs are properly substituted by other TAZs that have similar characteristics to the excluded TAZs. Comparing predictions of different local fitted models revealed a high predictive accuracy, substantiating the robustness of models.

5. CONCLUSIONS AND DISCUSSION

Application of Generalized Linear Models (GLM) with the assumption of Negative Binomial error distribution might be the most popular technique in crash prediction analysis. The results of GLM models are a set of fixed coefficient estimates which represent the average relationship between the dependent variable and other explanatory variables for all locations. These relationships are assumed to be constant across space. However, these explanatory variables are often found to be spatially heterogeneous especially when the study area is large enough to cover different traffic volume, urbanization and socio-demographic patterns. In this study we first aim to investigate the presence of spatial variation of dependent and different explanatory variables which are being used in developing crash prediction models. This was carried out by computing Moran's *I* statistics for dependent and selected explanatory variables. The results revealed the necessity of considering spatial correlation when developing crash prediction models. Therefore, different Geographically Weighted Poisson Regression (GWPR) models were developed, using different exposure, network and socio-demographic variables. GWPR models allow the estimations to vary where different spatial correlation among the variables exists. Hence, the association between NOICs and other explanatory variables are formed by means of different local models for each TAZ. Comparing models by means of MSPE and PCC show that local GWPR models always overperform global GLM models, both in fitting the data and predicting the response variable. This is due to the fact that GWPR models are capable of capturing the spatial heterogeneity of crash occurrence. Moreover, global estimates are unlikely to predict local changes properly. For planning at local levels (e.g. municipality level), local GWR models seem to be more appropriate, since global models might fail in capturing local changes. Furthermore, global models' predictions are more likely to be under/over estimated.

In construction of GWPR models different actions need to be taken. An important task is computing the most proper bandwidth and selecting the most suitable kernel function. For the current data, adaptive bandwidth with Gaussian kernel function result in the best model fit. Furthermore, the AICc method is adopted to compute bandwidth. This method relies on producing minimum AICc measure and has advantages compared to cross-validation (CV) method. Applying the CV method might increase the risk of over-fitting the calibration data, while the AICc method which penalizes possible small sample bias, accounts for the over-fitting issue.

Another issue that needs further discussion is the choice of the Poisson error distribution in this study. In traffic safety literature, utilizing the Negative Binomial error distribution is more favorable than the Poisson error distribution since it accounts for overdispersion that is commonly observed in crash data. However, since we accounted for spatial correlation in our models, it is expect that variance will become much closer to the mean (i.e. local models are fitted using a number of vicinity observation that are similar in their characteristics. This is demonstrated by means of Moran's *I* test for the number of crashes, indicating a significant clustering pattern.). This justifies the choice of Poisson error distribution that is adopted in this study.

REFERENCES

[1] N. Levine, K. E. Kim, and L. H. Nitz, “Spatial analysis of Honolulu motor vehicle crashes: I. Spatial patterns,” *Accident Analysis & Prevention*, vol. 27, no. 5, pp. 663–674, Oct. 1995.

[2] E. Amoros, J. L. Martin, and B. Laumon, “Comparison of road crashes incidence and severity between some French counties,” *Accident Analysis & Prevention*, vol. 35, no. 4, pp. 537–547, Jul. 2003.

[3] A. Hadayeghi, A. Shalaby, and B. Persaud, “Macrolevel Accident Prediction Models for Evaluating Safety of Urban Transportation Systems,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1840, no. -1, pp. 87–95, Jan. 2003.

[4] R. B. Noland and L. Oh, “The effect of infrastructure and demographic change on traffic-related fatalities and crashes: a case study of Illinois county-level data,” *Accident Analysis & Prevention*, vol. 36, no. 4, pp. 525–532, Jul. 2004.

[5] R. B. Noland and M. A. Quddus, “A spatially disaggregate analysis of road casualties in England,” *Accident Analysis & Prevention*, vol. 36, no. 6, pp. 973–984, Nov. 2004.

[6] F. L. D. De Guevara, S. Washington, and J. Oh, “Forecasting Crashes at the Planning Level: Simultaneous Negative Binomial Crash Model Applied in Tucson, Arizona,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1897, no. -1, pp. 191–199, Jan. 2004.

[7] G. R. Lovegrove, “Community-Based, Macro-Level Collision Prediction Models,” University of British Columbia, University of British Columbia, 2005.

[8] A. Hadayeghi, A. S. Shalaby, B. N. Persaud, and C. Cheung, “Temporal transferability and updating of zonal level accident prediction models,” *Accident Analysis & Prevention*, vol. 38, no. 3, pp. 579–589, May 2006.

[9] J. Aguero-Valverde and P. P. Jovanis, “Spatial analysis of fatal and injury crashes in Pennsylvania,” *Accident Analysis & Prevention*, vol. 38, no. 3, pp. 618–625, May 2006.

[10] G. R. Lovegrove and T. Sayed, “Macro-level collision prediction models for evaluating neighbourhood traffic safety,” *Canadian Journal of Civil Engineering*, vol. 33, no. 5, pp. 609–621, May 2006.

[11] A. Hadayeghi, A. S. Shalaby, and B. N. Persaud, “Development of planning level transportation safety tools using Geographically Weighted Poisson Regression,” *Accident Analysis & Prevention*, vol. 42, no. 2, pp. 676–688, Mar. 2010.

[12] G. Lovegrove and T. Sayed, “Macrolevel Collision Prediction Models to Enhance Traditional Reactive Road Safety Improvement Programs,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2019, no. -1, pp. 65–73, Dec. 2007.

[13] G. R. Lovegrove and T. Litman, “Using Macro-Level Collision Prediction Models to Evaluate the Road Safety Effects of Mobility Management Strategies: New Empirical Tools to Promote Sustainable Development,” presented at the Transportation Research Board (TRB) 87th Annual Meeting, Washington D.C. USA, 2008.

[14] A. Hadayeghi, “Use of Advanced Techniques to Estimate Zonal Level Safety Planning Models and Examine Their Temporal Transferability,” PhD thesis, Department of Civil Engineering, University of Toronto, PhD thesis, Department of Civil Engineering, University of Toronto, 2009.

[15] A. Naderan and J. Shahi, “Aggregate crash prediction models: Introducing crash generation concept,” *Accident Analysis & Prevention*, vol. 42, no. 1, pp. 339–346, Jan. 2010.

[16] D. Lord and F. Mannering, “The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives,” *Transportation Research Part A: Policy and Practice*, vol. 44, no. 5, pp. 291–305, Jun. 2010.

[17] M. An, C. Casper, and W. Wu, “Using Travel Demand Model and Zonal Safety Planning Model for Safety Benefit Estimation in Project Evaluation,” presented at the Transportation Research Board (TRB) 90th Annual Meeting, Washington D.C. USA, 2011.

[18] M. Abdel-Aty, C. Siddiqui, and H. Huang, “Integrating Trip and Roadway Characteristics in Managing Safety at Traffic Analysis Zones,” presented at the Transportation Research Board (TRB) 90th Annual Meeting, Washington D.C. USA, 2011.
[19] A. Hadayeghi, A. Shalaby, and B. Persaud, “Safety Prediction Models: Proactive Tool for Safety Evaluation in Urban Transportation Planning Applications,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2019, no. -1, pp. 225–236, Dec. 2007.
[20] N. Levine, K. E. Kim, and L. H. Nitz, “Spatial analysis of Honolulu motor vehicle crashes: II. Zonal generators,” *Accident Analysis & Prevention*, vol. 27, no. 5, pp. 675–685, Oct. 1995.
[21] E. A. LaScala, D. Gerber, and P. J. Gruenewald, “Demographic and environmental correlates of pedestrian injury collisions: a spatial analysis,” *Accident Analysis & Prevention*, vol. 32, no. 5, pp. 651–658, Sep. 2000.
[22] B. Flahaut, M. Mouchart, E. S. Martin, and I. Thomas, “The local spatial autocorrelation and the kernel method for identifying black zones: A comparative approach,” *Accident Analysis & Prevention*, vol. 35, no. 6, pp. 991–1004, Nov. 2003.
[23] E. Moons, T. Brijs, and G. Wets, “Identifying Hazardous Road Locations: Hot Spots versus Hot Zones,” presented at the International Conference on Computational Science and Its Applications (ICCSA), Perugia, Italy, 2009.
[24] B. Flahaut, “Impact of infrastructure and local environment on road unsafety: Logistic modeling with spatial autocorrelation,” *Accident Analysis & Prevention*, vol. 36, no. 6, pp. 1055–1066, Nov. 2004.
[25] K. Geurts, I. Thomas, and G. Wets, “Understanding spatial concentrations of road accidents using frequent item sets,” *Accident Analysis & Prevention*, vol. 37, no. 4, pp. 787–799, Jul. 2005.
[26] J. Aguero-Valverde and P. P. Jovanis, “Analysis of Road Crash Frequency with Spatial Models,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2061, no. -1, pp. 55–63, Dec. 2008.
[27] C. D. Cottrill and P. (Vonu) Thakuriah, “Evaluating pedestrian crashes in areas with high low-income or minority populations,” *Accident Analysis & Prevention*, vol. 42, no. 6, pp. 1718–1728, Nov. 2010.
[28] H. Huang, M. Abdel-Aty, and A. Darwiche, “County-Level Crash Risk Analysis in Florida,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2148, no. -1, pp. 27–37, Dec. 2010.
[29] S.-P. Miaou, J. J. Song, and B. K. Mallick, “Roadway Traffic Crash Mapping: A Space-Time Modeling Approach,” *Journal of Transportation and Statistics*, vol. 6, no. 1, pp. 33–57, 2003.
[30] X. Wang and M. Abdel-Aty, “Temporal and spatial analyses of rear-end crashes at signalized intersections,” *Accident Analysis & Prevention*, vol. 38, no. 6, pp. 1137–1150, Nov. 2006.
[31] M. A. Quddus, “Modelling area-wide count outcomes with spatial correlation and heterogeneity: An analysis of London crash data,” *Accident Analysis & Prevention*, vol. 40, no. 4, pp. 1486–1497, Jul. 2008.
[32] C. Wang, M. A. Quddus, and S. G. Ison, “Impact of traffic congestion on road accidents: A spatial analysis of the M25 motorway in England,” *Accident Analysis & Prevention*, vol. 41, no. 4, pp. 798–808, Jul. 2009.
[33] F. Guo, X. Wang, and M. Abdel-Aty, “Modeling signalized intersection safety with corridor-level spatial correlations,” *Accident Analysis & Prevention*, vol. 42, no. 1, pp. 84–92, Jan. 2010.
[34] C. Siddiqui, M. Abdel-Aty, and K. Choi, “Macroscopic spatial analysis of pedestrian and bicycle crashes,” *Accident Analysis & Prevention*, vol. 45, no. 0, pp. 382–391, Mar. 2012.
[35] A. S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression the analysis of spatially varying relationships*. West Sussex, England: John Wiley & Sons Ltd, 2002.
[36] F. Zhao and N. Park, “Using Geographically Weighted Regression Models to Estimate Annual Average Daily Traffic,” *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1879, no. -1, pp. 99–107, Jan. 2004.

[37] A. Páez, "Exploring contextual variations in land use and transport analysis using a probit model with geographical weights," *Journal of Transport Geography*, vol. 14, no. 3, pp. 167–176, May 2006.
[38] L.-F. Chow, F. Zhao, X. Liu, M.-T. Li, and I. Ubaka, "Transit Ridership Model Based on Geographically Weighted Regression," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1972, no. -1, pp. 105–114, Jan. 2006.
[39] H. Du and C. Mulley, "Relationship Between Transport Accessibility and Land Value: Local Model Approach with Geographically Weighted Regression," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 1977, no. -1, pp. 197–205, Jan. 2006.
[40] S. D. Clark, "Estimating local car ownership models," *Journal of Transport Geography*, vol. 15, no. 3, pp. 184–197, May 2007.
[41] S. Blainey, "Trip end models of local rail demand in England and Wales," *Journal of Transport Geography*, vol. 18, no. 1, pp. 153–165, Jan. 2010.
[42] E. C. Delmelle and J.-C. Thill, "Urban Bicyclists: Spatial Analysis of Adult and Youth Traffic Hazard Intensity," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2074, no. -1, pp. 31–39, Dec. 2008.
[43] S. Erdogan, "Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey," *Journal of Safety Research*, vol. 40, no. 5, pp. 341–351, Oct. 2009.
[44] A. Pirdavani, T. Brijs, T. Bellemans, B. Kochan, and G. Wets, "Developing Zonal Crash Prediction Models with a Focus on Application of Different Exposure Measures," *Transportation Research Record: Journal of the Transportation Research Board*, 2012.
[45] D. Janssens, G. Wets, H. J. P. Timmermans, and T. A. Arentze, "Modelling Short-Term Dynamics in Activity-Travel Patterns: Conceptual Framework of the Feathers Model," presented at the 11th World Conference on Transport Research, Berkeley CA, USA, 2007.
[46] B. Kochan, T. Bellemans, D. Janssens, and G. Wets, "Assessing the Impact of Fuel Cost on Traffic Demand in Flanders Using Activity-Based Models," presented at the Travel Demand Management TDM, Vienna, Austria, 2008.
[47] L. Guo, Z. Ma, and L. Zhang, "Comparison of bandwidth selection in application of geographically weighted regression: a case study," *Canadian Journal of Forest Research*, vol. 38, no. 9, pp. 2526–2534, Sep. 2008.
[48] T. Nakaya, A. S. Fotheringham, C. Brunsdon, and M. Charlton, "Geographically weighted Poisson regression for disease association mapping," *Statistics in Medicine*, vol. 24, no. 17, pp. 2695–2717, Sep. 2005.
[49] *R: A language and environment for statistical computing*. Vienna, Austria: R Development Core Team, 2011.
[50] V. Y.-J. Chen and T.-C. Yang, "SAS macro programs for geographically weighted generalized linear modeling with spatial point data: Applications to health research," *Computer Methods and Programs in Biomedicine*, vol. 107, no. 2, pp. 262–273, Aug. 2012.
[51] D. Wheeler and M. Tiefelsdorf, "Multicollinearity and correlation among local regression coefficients in geographically weighted regression," *Journal of Geographical Systems*, vol. 7, no. 2, pp. 161–187, 2005.
[52] M. H. Kutner, C. J. Nachtsheim, and J. Neter, *Applied Linear Regression Models*, 4th ed. McGraw-Hill, 2004.